

ON RECOVERY OF SPARSE SIGNALS IN COMPRESSED DNA MICROARRAYS

H. Vikalo¹, F. Parvaresh², and B. Hassibi²

¹Department of Electrical and Computer Engineering, University of Texas, Austin, TX

²Department of Electrical Engineering, California Institute of Technology, Pasadena, CA

ABSTRACT

Currently, DNA microarrays comprising tens of thousands of probe spots are employed to test entire genomes in a single experiment. Typically, each microarray spot contains a large number of copies of a single probe, and hence collects only a single data point. This is a wasteful use of the sensing resources in comparative DNA microarray experiments, where a test sample is measured relative to a reference sample. Since only a small fraction of the total number of genes represented by the two samples is differentially expressed, a large fraction of a microarray does not provide any useful information. To this end, in this paper we consider an alternative microarray design wherein each spot is a composite of several different probes, and the total number of spots is potentially much smaller than the number of genes being tested. Fewer spots directly translates to significantly lower costs due to cheaper array manufacturing, simpler image acquisition and processing, and smaller amount of genomic material needed for experiments. To recover signals from compressed microarray measurements, we leverage ideas from compressive sampling. Experimental verification of the proposed methodology is presented.

1. INTRODUCTION

Sensing in DNA microarrays [1] is based on the process of hybridization in which complementary DNA strands bind to each other creating structures in lower energy states. Typically, the surface of a DNA microarray comprises an array of spots, each spot containing a large number of identical single-stranded DNA sequences (*probes*) designed to capture copies of a single DNA molecule (*target*) of interest. DNA microarrays are often used to measure gene expression levels, i.e., to quantify the process of transcription of DNA information into messenger RNA molecules. The information transcribed into mRNA is further translated to proteins, the molecules that perform most of the functions in cells. Therefore, by measuring gene expression levels, we may be able to infer critical information about the functionality of cells or whole organisms [2], study diseases and the effects of drugs on

them [3, 4], etc. DNA microarrays are often used to compare the gene expression levels of a test sample with that of a reference sample. In a typical scenario, only a small fraction of the total number of genes is differentially expressed. For instance, only several hundreds genes (out of, say, 30,000 in an entire genome), may be differentially expressed. Therefore, a large fraction of a microarray does not contribute any information about the subset of the genes that are differentially expressed. To remedy this, in [5] a microarray architecture comprising spots that contain mixtures of several different probes was proposed, so that a signal measured at each probe spot is potentially a combination of as many targets. This allows acquisition of multiple data points for each of the targets being tested, including those that are indeed differentially expressed. However, the signal recovery in the *composite microarrays* [5] does not exploit sparseness of the signal.

By leveraging ideas from *compressive sampling*, we can enable more economic usage of the sensing resources in composite microarrays. The essential idea of compressive sampling is that we may be able to recover an inherently sparse signal by using far fewer measurements than what is typically needed for a signal which is not sparse [6]. Compressive sampling is closely related to the problem of solving an underdetermined system of linear equation with a sparseness constraint – which is precisely the problem of signal recovery in composite microarrays with fewer probe spots than probes. In fact, by judiciously choosing probes comprising each spot, we may be able to recover sparse signal from a microarray wherein the number of probe spots is significantly reduced. We refer to such platforms as *compressed microarrays*. Having fewer probe spots translates to lower costs due to cheaper array manufacturing, simpler image acquisition and processing, and smaller amount of genomic material needed for experiments. Moreover, decreasing sample volume size is critically important in order to further the applications of microarray technology in diagnostics and environmental monitoring applications.

Typically, DNA microarrays are manufactured by either spotting (i.e., printing) probe molecules in their allotted spots, or by a direct probe synthesis on the array. While the former technique can directly be applied to manufacturing compressed microarrays (by, e.g., spotting appropriately selected mixtures of probes), it is not imme-

This work was supported in part by a Grubstake Award from California Institute of Technology, a grant from the David and Lucille Packard Foundation, and by the Millard and Muriel Jacobs Genetics and Genomics Laboratory at Caltech.

diately clear how the latter could be done. In the current work, we focus on the former manufacturing technique, i.e., we design, analyze, and experiment with the compressed microarrays manufactured by probe spotting.

2. BACKGROUND

In compressive sampling, we are interested in estimating an n -dimensional signal \mathbf{x} which has no more than k non-zero entries. (Note that we do not know a priori the locations of the non-zero entries.) So, $k < n$; in fact, we frequently focus on applications where $k \ll n$.

The vector \mathbf{x} is not directly observable. Instead, we observe m linear combinations of the entries of \mathbf{x} ,

$$y_i = \sum_{j=1}^n A_{ij} x_j, \quad i = 1, 2, \dots, m, \quad (1)$$

where $k < m < n$. In other words, the number of measurements that we collect is smaller than the size of the vector \mathbf{x} , yet larger than the number of its non-zero entries. Collecting the coefficients A_{ij} into an $m \times n$ matrix A , we can write (1) in a matrix form

$$\mathbf{y} = A\mathbf{x}. \quad (2)$$

The underdetermined system of equations (2) may, in principle, be solved by using the fact that the vector \mathbf{x} is sparse. In particular, we could consider all possible combinations of k columns of A , and attempt to solve the corresponding system of equations which is overdetermined (since each one has m equations with k unknowns). Assuming that each of these combinations of columns forms a matrix with a full rank, at least one of the overdetermined systems will have a solution. This solution determines the positions and values of the non-zero entries in \mathbf{x} . However, the outlined approach is clearly practically infeasible.

On the other hand, for a long time it has been known that constrained l_1 minimization,

$$\min_{\mathbf{x}, A\mathbf{x}=\mathbf{y}} \|\mathbf{x}\|_1, \quad (3)$$

as well as the related constrained quadratic programming

$$\min \|y - Ax\|_2 \text{ subject to } \|x\|_1 \leq \beta, \quad (4)$$

where $\|\mathbf{x}\|_1 = \sum_{i=1}^n |x_i|$ denotes the l_1 -norm of the vector \mathbf{x} , and β is an appropriately chosen constant, perform well when employed for finding sparse solutions (see, e.g., [9]). Only recently there have been theoretical results justifying the performance of the constrained l_1 minimization. These results show that, for measurement matrices A which satisfy certain conditions, the constrained l_1 minimization recovers the solution if the unknown vector \mathbf{x} is sparse enough, i.e., if the ratio k/n is sufficiently small [7].

Finally, we should mention that, in the course of preparation of the current paper, we became aware of the related work [11], which also proposes the use of compressed sensing techniques. However, unlike our method which involves printing several different probe types in each spot of the microarray (and therefore leads to a sparse measurement matrix – see the section below), [11] proposes the design of probes, each of which can potentially capture several different targets. Calibrating such an array (in the sense of determining the strength of the binding of each target analyte to its corresponding probe) can be challenging. Our approach, however, can use already-designed probe sets and simply requires mixing a number of them prior to spotting them on the array – a procedure which is readily feasible.

3. COMPRESSED MICROARRAYS

When quantifying a sparse signal, compressive sampling provides cost-efficient utilization of the sensing resources. In particular, recall that a sparse signal may be recovered from a small number of linear combinations of its components. The compressive sampling ideas are relevant to the applications of DNA microarrays in gene expression profiling, where the gene expression levels of a test sample are compared with the gene expression levels of a reference sample. Since in practical scenarios only a small fraction of the total number of genes is differentially expressed, the difference of the signals produced by the two samples is sparse. Moreover, linear combinations of the signal components may be acquired by the composite probe spots comprising a mixture of several probe sequences as in [5]. The sparseness constraint, on the other hand, suggests possible recovery of the signal from potentially far fewer probe spots than the total number of probe sequences composing the spots of the microarray.

In [12], we developed a statistical model for microarrays, which is directly applicable to the compressed microarrays. In particular, for a compressed microarray with n spots containing probes designed to quantify m different targets, we can write

$$\mathbf{y} = A\mathbf{x} + \mathbf{w} + \mathbf{v}, \quad (5)$$

where \mathbf{y} denotes the n -dimensional measurement, \mathbf{x} denotes the m -dimensional data vector (the number of copies of each target), \mathbf{v} is the n -dimensional zero-mean iid Gaussian additive noise due to instrumentation and other biochemistry-independent noise sources, \mathbf{w} denotes the shot-noise (i.e., zero-mean iid Gaussian noise with covariance proportional to the signal – see, e.g., [12]), and where A is an $n \times m$ binary matrix containing information about probe mixing. In other words, the (i, j) element of A is non-zero if and only if the j th target can bind to some of the probes in the i th spot. We limit the entries in A to binary 1/0 for the

sake of manufacturing simplicity, e.g., to impose the constraint that each microarray spot contains an equal amount of different probes comprising it. Each row of the matrix A corresponds to a probe spot. The composition of the i^{th} probe spot, $1 \leq i \leq m$, is determined by the positions of ones in the i^{th} row of A . Moreover, the number of different probes in the i^{th} spot is equal to the number of ones in the i^{th} row of the matrix A .

In a two-color microarray experiment, we are comparing two samples characterized by data vectors \mathbf{x}_1 and \mathbf{x}_2 , and are interested in finding differentially expressed genes, i.e., finding non-zero entries of the vector $x = \mathbf{x}_1 - \mathbf{x}_2$. Defining $y = \mathbf{y}_1 - \mathbf{y}_2$, $w = \mathbf{w}_1 - \mathbf{w}_2$, and $v = \mathbf{v}_1 - \mathbf{v}_2$, we can write

$$y = Ax + w + v. \quad (6)$$

The vector x in (6) is sparse, i.e., it has a small number of entries that are non-zero (or significantly larger than zero). Recalling the discussion of compressive sampling, it should appear clear that since x is sparse, one may be able to recover it using (3) or (4).

We should briefly mention the important issue of probe design. Two among the most important properties of microarray probes are their sensitivity and specificity. Sensitivity is a measure of how strongly a probe reacts with the target which it is supposed to capture. Specificity, on the other hand, is the ability of a probe to discriminate between targets, i.e., its ability to ignore (do not bind or cross-hybridize to) other targets. In (6), we have implicitly assumed that all probes are equally sensitive and that there is no probe-target binding due to cross-hybridization. The scenario wherein these assumptions do not hold and techniques which take that into account are considered in [12]. Imbalanced sensitivity, for instance, may be incorporated in the compressed microarray model by appropriately scaling selected non-zero entries of A . Imperfect specificity, on the other hand, would require increasing the fraction of non-zero entries in A . In general, cross-hybridization is detrimental to the complexity of the signal recovery in compressed microarrays and thus special attention should be paid to specificity of probes in compressed microarrays.

As an illustration, in Figure 1 we demonstrate the performance of l_1 -constrained minimization employed for the detection of sparse signals in a compressed microarray simulated according to the model (6). The microarray comprises $n = 24$ probe spots, and each spot contains a mixture of 24 different probes chosen from the set of $m = 96$ available probe sequences, each designed to capture one target of interest. So, the dimension of the matrix A is 24×96 . Moreover, the number of non-zero entries in x is $k = 8$. Parameters of the microarray model (6) are chosen so as to mimic a realistic experiment. As implied by Figure 1, the algorithm successfully recovers sparse

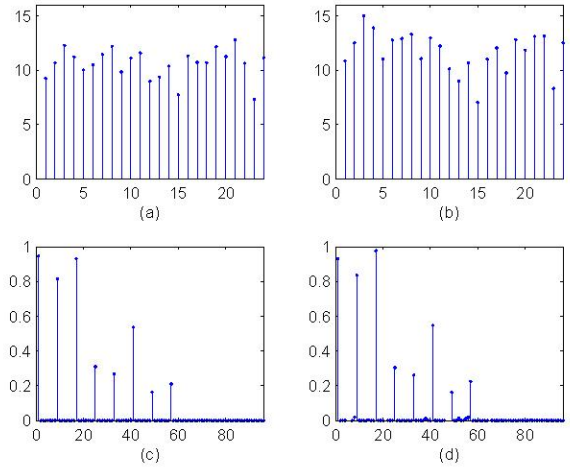


Fig. 1. Demonstration of the sparse signal recovery in a compressed microarray. Subfigures (a) and (b) show the test and the reference signals, respectively, versus probe spot index. Subfigure (c) shows the sparse signal, and subfigure (d) its estimate obtained by solving an appropriate l_1 minimization problem.

data from noisy observations.

3.1. Compressed microarrays for aCGH technologies

Array comparative genomic hybridization (aCGH) technology, wherein total genomic DNA from a test and a reference samples are compared on a DNA microarray, has recently emerged as a platform for detecting and mapping alterations in genomic structures. The arrays used in aCGH applications may have as many as million of distinct probes and are thus capable of providing high-resolution information about the DNA sequence in focus. The aforementioned alterations (i.e., changes in DNA copy number) include small amplifications and deletions, as well as large chromosomal gains and losses [13]. Such alterations are commonly encountered in various types of human cancers, including breast and ovarian (see, e.g., [13] and the references therein). Therefore, detecting DNA alterations can provide valuable information about the genomic mechanisms of cancer as well as be used in diagnostics and search for drug treatments.

The alterations typically affect continuous segments of a genome, and their total length is just a fraction of the length of entire genome. By mapping DNA segments to their chromosomal locations and reordering x accordingly, the unknown signal is both sparse and piecewise constant (see, e.g., [13] and the references therein). Sparse signal with piecewise constant segments can be efficiently recovered using l_1 minimization. For instance, the Lasso

optimization problem (4) can be modified to

$$\min \|y - Ax\|_2 \text{ subject to } \|x\|_1 \leq \beta, \|Dx\|_1 \leq \gamma, \quad (7)$$

where β and γ are appropriately chosen constants, and where the matrix D denotes the differentiation operator, i.e., its general structure is given by

$$D = \begin{bmatrix} 1 & -1 & 0 & 0 & 0 \\ 0 & 1 & -1 & 0 & 0 \\ & & \ddots & \ddots & \\ 0 & 0 & 0 & 1 & -1 \end{bmatrix}.$$

[We should note that other constraints on the signal, such as piecewise linear, can be taken into account in a way similar to (7).]

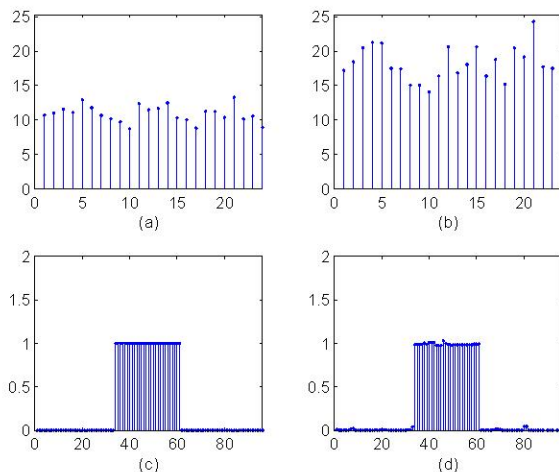


Fig. 2. Illustration of the sparse signal recovery in a compressed microarray for aCGH applications. The toy example considers an array with $n = 24$ probe spots, where each spot comprises mixtures selected from the set of $m = 96$ probes. The number of non-zero entries in the 96-dimensional piecewise-constant signal is $k = 28$. Subfigures (a) and (b) show the test and the reference signals, respectively, versus probe spot index. Subfigure (c) shows the sparse signal, and subfigure (d) its estimate obtained by solving (7).

In Figure 2, we illustrate the performance of the algorithm (7) when applied for detection of sparse, piecewise-constant signals in compressed microarrays simulated according to the model (6). The microarray comprises $n = 24$ probe spots, and each spot contains a mixture of 24 different probes chosen from the set of $m = 96$ available probe sequences, each designed to capture one target of interest. Hence, the dimension of the matrix A is 24×96 . Moreover, the number of the non-zero entries in

is $k = 28$. The simulation parameters of the model (6) are chosen so as to mimic realistic experiment. As implied by Figure 2, (7) successfully recovers sparse data from noisy observations. Note that the fraction of non-zero components of x is fairly large ($k = 28$ out of $m = 96$). In fact, without the additional piecewise-constant constraint, l_1 minimization is not able to recover the signal since it is not sufficiently sparse. However, since the derivative of a piecewise-constant signal is very sparse, the signal can be efficiently recovered using (7).

4. EXPERIMENTAL VERIFICATION

In this section, we present several proof-of concept experiments designed and conducted to demonstrate data acquisition and signal recovery in compressed microarrays. The goal was detection and quantification of $k \leq 8$ targets on an array otherwise capable of testing $n = 96$ different targets. The desired probe spot compression ratio, m/n was chosen to be 4. Therefore, the compressed microarray has only $m = 24$ probe spots, each comprising a combination of a number of different probe sequences. Mixtures of the probes, synthesized oligonucleotide sequences, were deposited to their respective spots; the targets are cDNA molecules extracted from *Escherichia Coli*. In particular, the targets were generated using The RNA Spikes™, a commercially available set of 8 purified RNA transcripts purchased from Ambion Inc. Typically, these spikes are used in microarrays for calibration purposes and have been chosen so that the eight sequences have little mutual correlation. The RNA sequences were reverse transcribed to obtain cDNA targets, which were then labeled with Cy5 dyes. We denote the set of these 8 targets by T_8 .

Eight oligo probes designed for capturing the targets in T_8 were also purchased from Ambion Inc. Moreover, we acquired 88 probes designed to test the mouse genome. We denote the set of Ambion probes as P_8 , and the set of mouse genome probes as P_{88} . The full set of 96 oligonucleotide probes, all of them 25 nucleotides long, is denoted as P_{96} . The targets from T_8 do not cross-hybridize with (i.e., bind to) the probes from P_{88} . We designed $m = 24$ different mixtures, each comprising 24 probes selected from P_{96} . Each of the mixtures is deposited in one of the spots of the compressed microarray. Content of the mixtures determine composition of the coefficient matrix A ; hence, each row in A has 24 ones and 72 zeros.

The sparse signal vector x was constructed such that $x_k \neq 0$ if and only if $k \in \mathcal{K} = \{1, 9, 17, 25, 33, 41, 49, 57\}$. In particular, x_1 contains information about the amount of the first target from the set T_8 , x_9 contains information about the amount of the second target from T_8 , etc. The targets from T_8 were applied to a microarray, where the individual amounts of targets were (5ng, 5ng, 2ng, 1ng,

10ng, 2ng, 1ng, 1ng), respectively. The experiment was run overnight and the array, after washing away the sample, was scanned. Figure 5 shows (a) the measured light intensities of the compressed microarray spots, and (b) the recovered signal. Clearly, the strongest 8 components of the recovered signal correspond to the targets in T_8 .

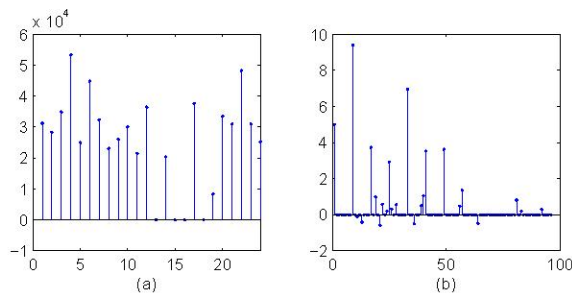


Fig. 3. Measured (a) and recovered (b) signal.

We conducted several more compressed microarray experiments testing the targets from T_8 , sometimes adding complex biological background (i.e., total mice DNA) to the sample; in these experiments, the strong components of the recovered signal vector correctly identified targets from T_8 and thus the compressed microarray proved capable of detecting their presence. As a part of the future work, we intend to calibrate the array (i.e., determine the affinities of the targets from T_8 to their corresponding probes) in order to enable precise quantification of their amounts.

5. SUMMARY AND CONCLUSIONS

We presented a novel DNA microarray architecture which we refer to as compressed DNA microarrays. In compressed microarrays, each probe spot contains a mixture of a number of different probes. By exploiting inherent sparseness of the signals in gene expression studies, target detection and quantification can be performed on an array with significantly reduced number of spots. To this end, we used ideas from compressive sampling, and employed linear programming to solve an appropriate l_1 -minimization problem. Both simulations as well as experiments confirm that if the signal vector is sufficiently sparse, l_1 -minimization can recover it.

There are many directions where the work presented in the current paper can be extended. There is a need to find deterministic coefficient matrices that are sparse and have the properties required for precise signal recovery.

6. REFERENCES

- [1] M. Schena et. al., "Quantitative monitoring of gene expression patterns with a complementary DNA mi-

croarray," *Science*, 270(5235), October 1995, pp. 467-70.

- [2] M. Schena et. al., "Microarrays: biotechnology's discovery platform for functional genomics," *Trends in Biotechnology* 1998, 16, 301-306.
- [3] J. Kononen et. al., "Tissue microarrays for high-throughput molecular profiling of tumor specimens," *Nature Medicine*, 4(7), July 1998, pp. 844-847.
- [4] D. T. Ross et. al., "Systematic variation in gene expression patterns in human cancer cell lines," *Nature Genetics*, 24(3), March 2000, pp. 227-35.
- [5] I. Shmulevich, J. Astola, D. Cogdell, S. R. Hamilton, and W. Zhang, "Data extraction from composite oligonucleotide microarrays," *Nucleic Acids Research*, vol. 31, no. 7, 2003.
- [6] E. J. Candes, "Compressive sampling," *Proc. of the Intern. Congress of Mathem.*, Madrid, Spain, 2006.
- [7] E. Candes and T. Tao, "Decoding by linear programming," *IEEE Transactions on Information Theory*, 51(12), December 2005, pp. 4203-4215.
- [8] E. Candes, J. Romberg, and T. Tao, "Stable signal recovery from incomplete and inaccurate measurements," *Comm. on Pure and Appl. Math.*, 59(8), August 2006, pp. 1207-1223.
- [9] M. J. Wainwright, "Sharp thresholds for high-dimensional and noisy recovery of sparsity," *Proc. 44th Ann. Allerton Conf. on Comm., Contr., and Comput.*, Monticello, IL, 2006.
- [10] E. Candes and J. Romberg, "Quantitative robust uncertainty principles and optimally sparse decompositions," *Foundations of Comput. Math.*, 6(2), April 2006, pp. 227-254.
- [11] O. Milenkovic, R. Baraniuk, and T. Simunic-Rosing, "Compressed sensing meets bionformatics: a new DNA microarray architecture," *Inform. Theory and Applications Workshop*, San Diego, 2007.
- [12] H. Vikalo, A. Hassibi, and B. Hassibi, "A statistical model for microarrays, optimal estimation algorithms, and limits of performance," *IEEE Trans. on Sig. Proc., Spec. Issue on Gen. Sig. Proc.*, vol. 54, no. 6, June 2006.
- [13] R. Pique-Regi, E.-S. Tsau, A. Ortega, R. Seeger, and S. Asgharzadeh, "Wavelet footprints and sparse Bayesian learning from DNA copy number change analysis," *Proc. of the IEEE Intern. Conf. on Acoust., Speech, and Sig. Proc.*, 2007, pp. 353-356.